Risks and benefits of speech recognition for clinical documentation: a systematic review

RECEIVED 23 May 2015
REVISED 31 August 2015
ACCEPTED 4 September 2015
PUBLISHED ONLINE FIRST 17 November 2015

Tobias Hodgson and Enrico Coiera

ABSTRACT

Objective To review literature assessing the impact of speech recognition (SR) on clinical documentation.
Methods Studies published prior to December 2014 reporting clinical documentation using SR were identified by searching Scopus, Compendex and Inspect, PubMed, and Google Scholar. Outcome variables analyzed included dictation and editing time, document turnaround time (TAT), SR accuracy, error rates per document, and economic benefit. Twenty-three articles met inclusion criteria from a pool of 441.
Results Most studies compared SR to dictation and transcription (DT) in radiology, and heterogeneity across studies was high. Document editing time increased using SR compared to DT in four of six studies (+1876.47% to –16.50%). Dictation time similarly increased in three of five studies (+91.60% to –25.00%). TAT consistently improved using SR compared to DT (16.41% to 82.34%); across all studies the improvement was 0.90% per year. SR accuracy was reported in ten studies (88.90% to 96.00%) and appears to improve 0.03% per year as the technology matured. Mean number of errors per report increased using SR (0.05 to 6.66) compared to DT (0.02 to 0.40). Economic benefits were poorly reported.
Conclusions SR is steadily maturing and offers some advantages for clinical documentation. However, evidence supporting the use of SR is weak, and further investigation is required to assess the impact of SR on documentation error types, rates, and clinical outcomes.

INTRODUCTION

Speech recognition (SR) systems for medical reporting have been available commercially for over two decades.[1] SR is an input mechanism available to assist with clinical documentation by translating speech into text, or verbally controlling user interface functions. SR has been adopted successfully in clinical settings such as the dictation of radiology reports where it is used in conjunction with the radiology information system or picture archiving and communication systems.[2] However, SR has not been uniformly used across all clinical domains.[3] In contrast, SR is now widely used in many consumer applications, including interface control and question answering applications in smart phones.[4]

Early adoption of SR-based documentation was hindered by immature technology and clinically unacceptable recognition error rates,[5] but steady advances in recognition algorithm design and system performance have been made over the last twenty years.[6] In particular, the underlying technology within SR systems has evolved dramatically with advances in both the SR engines used to recognize speech, as well as the speed and memory of the hardware used to process speech data. Early SR systems utilized formal language grammars, but these were superseded by probabilistic approaches such as Hidden Markov models, which persist to the present day.[4,7] Such language models also require signal processing methods to extract basic features from speech data, and statistical acoustic models which represent the different sounds or phonemes in speech.[8,9] SR methods continue to evolve and newer methods include structured speech and language models,[10] conditional random fields, and maximum entropy Markov models.[11]

The objective of this review is to summarize the research literature describing the benefits and risks associated with the use of SR systems for clinical documentation tasks. A secondary aim is to explore whether SR performance in clinical documentation tasks has improved over time as this technology class has matured.

METHODS

A PRISMA compliant systematic review of studies examining the use of SR for clinical documentation was undertaken.[12]

REVIEWS

Correspondence to Tobias Hodgson, Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, L6 75 Talavera Rd, North Ryde, NSW 2109, Australia; tobias.hodgson@students.mq.edu.au; Tel: +61 2 9850 2425

For affiliation see end of article.

To be included in the review, studies needed to meet the following criteria:

- The article was published in English.
- Participants were clinicians performing clinical documentation tasks.
- The documentation intervention was SR.
- Quantitative outcomes were reported, which could include: experimental and observational study designs including randomized controlled trials, quasi-experimental studies, before-and-after studies, case-control studies, cohort studies, and cross-sectional studies.
- Measured outcomes were reported including one or more of: document turnaround time, error rates per document, dictation and editing time, SR accuracy, and/or economic benefit.

Abstracts in which full study data was unavailable were excluded. Study quality was assessed by examining study design, bias risk, duration, population size, reporting tasks, and number and type of errors reported.

Article searches were made using Scopus, Compendex and Inspect, PubMed, and Google Scholar with no date restriction. The search query used was: "speech recognition" or "voice recognition" or "Dragon Naturally Speaking" (the inclusion of other brands product names did not result in additional results) and "medical record*" or "health record*" or "patient record*" or "nursing record*" or "clinical record*" or "radiolog*". These searches identified 538 potential articles: Scopus (307), Compendex and Inspect (18), PubMed (9), and Google Scholar (204). Titles and abstracts were then identified and screened with 361 initial exclusions, 55 cases of being unable to obtain full text or requiring further information to make an assessment, leaving 122 full texts that were retrieved and evaluated. Each article was assessed independently by two reviewers (T.H. and D.A.) against the inclusion criteria. After assessment, twenty-three studies remained (Figure 1). In instances of disagreement (four articles), after deliberation, a consensus assignment was made (three of the four were excluded).

Study data including the intervention, population, study design, and outcomes were extracted using a standardized template (Appendix B, Table B1). Outcome variables were summarized but could not be pooled because of study heterogeneity. Mean, upper, and lower limits were documented and, where feasible, percentage change due to the intervention was calculated. Temporal trends were estimated using linear regression. Where confidence intervals were calculated, a normal distribution was assumed. Two reviewers also assessed the overall disposition of a study to the use of SR for clinical documentation as either positive or negative. Inter-rater agreement was calculated using Cohen's κ.[13] The observed agreement was statistically significant at 95.65% with a κ of 0.90.

RESULTS

There was heterogeneity in study design spanning randomized controlled trials, noncontrolled trials, and cross-sectional studies. There was also heterogeneity in clinical documentation task and settings. In the comparative effectiveness trials reported, SR was only compared to dictation and transcription (DT).[14] Most documentation tasks were radiological (fifteen of twenty-three reports), and the remainder were tasks from: pathology, endocrinology, dental, or general clinical documentation. The majority of clinical documentation tasks were free text based as opposed to templates/structured reporting (seventeen of twenty-three reports).

Efficiency of speech recognition

Many of the studies reported on the impact of SR on the time to create or modify a clinical document by a clinician, as well as the impact on organizational turnaround in document processing Table 1.

Five studies reported dictation time, which is the time taken to create a new clinical document using either SR or DT[14–18] Four of the five studies analyzed radiology reports and one studied general clinical notes.[18] Three showed an increase in dictation time using SR from 35.64%[15] to 91.60%.[16] Two studies showed a reduction in dictation time using SR compared to DT of 10.87%[17] and 25.00%.[18] Each study used a different SR system: Philips[14] Nuance,[18] LTI,[17] AGFA,[16] and ASR Medispeak.[15] Sample sizes ranged from one subject[18] to a whole department[17] and study duration ranged from 4 weeks[17] to 3 months.[16] Six studies compared the time required to edit documentation using SR compared to DT.[16–21] Editing involved review, alteration, or finalization of clinical documents in preparation for submission. Four studies reported a significant increase in editing time using SR to create reports[16,18,20,21] and two showed a small reduction.[17,19] Changes to editing time ranged from an increase of 1876.47% (17 s compared to 336 s using SR)[18] to a reduction of 16.50% (303 s down to 253 s).[19] Three studies examined radiology reports,[16,19,21] two looked at general clinical notes,[17,18] and one examined pathology reports.[20]

There were eight studies that compared turnaround time (TAT), which was defined as the time taken for the entire process from report creation to completion and submission[14,19,22,27] All showed a consistent decrease in TAT when using SR. There was a maximum of an 81.16% decrease (1486 min down to 280 min)[26] and a minimum decrease of 16.41% (329 min down to 275 min).[19] These studies were predominately free text reports[14,19,22,24,26] and seven of the eight were radiological[14,19,22–24,26,27] When analyzed over time, there appears to be a modest improvement in TAT across all studies of ~0.90% per year (Figure 2).

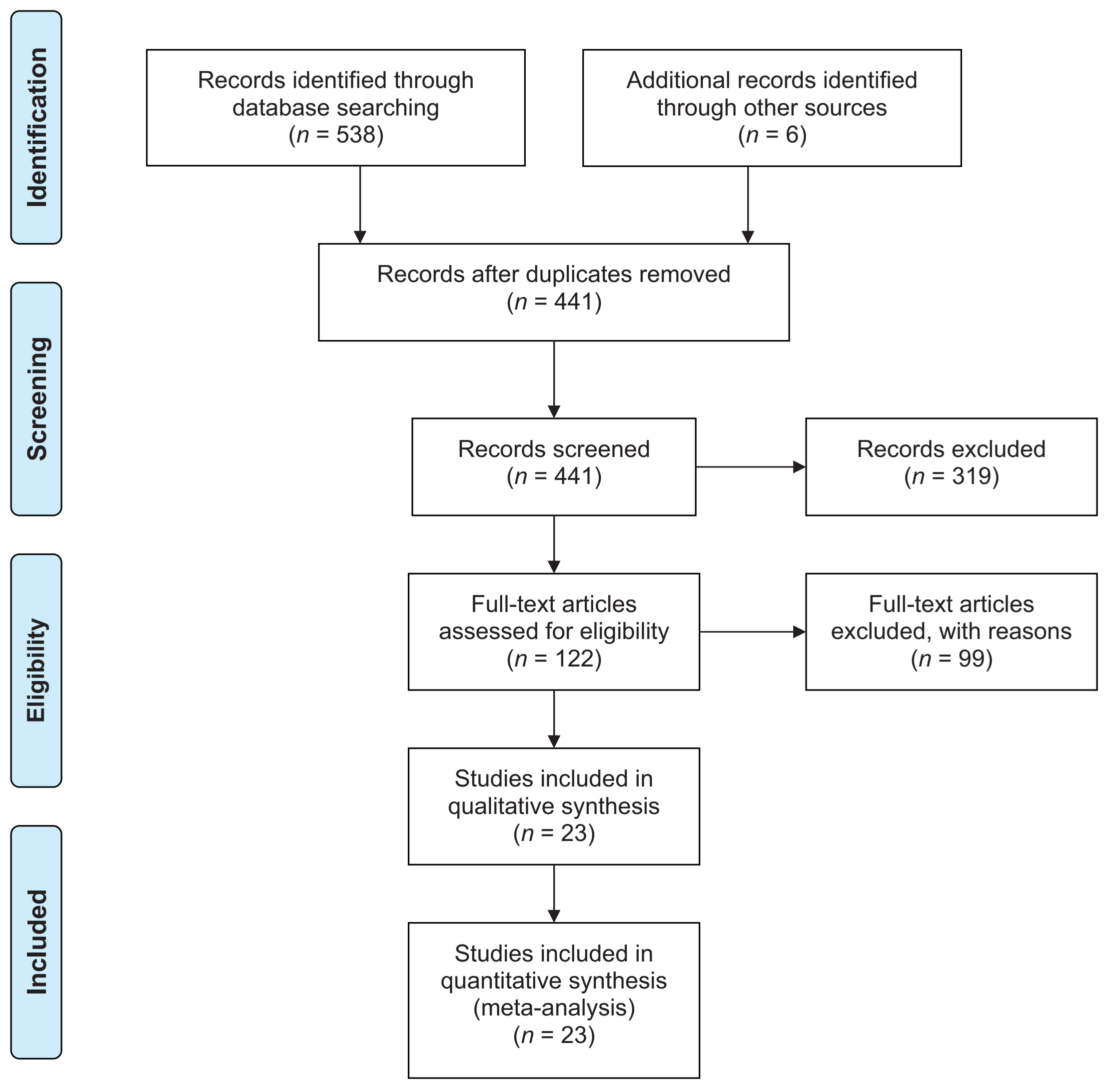
Four studies compared the number of words per report using SR or DT.[16,18,21,24] Three saw a reduction in the words per report when using SR ranging from 13.13%[18] to 36.84%.[24] One study[21] saw SR increase the number of words in a clinical document by 14.30%. Three were radiology studies[16,21,24] and three were free text reports.[16,21,24]

Accuracy of speech recognition

Several studies reported on the accuracy of SR as a data entry mechanism, reporting on variables including: number of errors per document (SR and DT) and the SR accuracy rate.

REVIEWS

Figure 1: Systematic review flow diagram based upon the PRISMA guidelines.[12]



Ten studies reported SR accuracy rates[14,16,18–20,24,25,28–30] ranging from a mean accuracy of 88.90%[1] to 96.00%.[16] Overall accuracy rates across all reports showed minimal improvement over time, at 0.03% per year (Figure 3). Six of the studies compared mean errors per document for both DT and SR.[15,16,18,20,31,32] All showed a substantial increase in the number of errors when using SR. Mean errors per document using DT varied from 0.02[31] to 0.40.[20] In contrast the mean errors per report created using SR was typically far higher varying from 0.05[31] to 6.66.[20] The number of mean additional errors found per report when created using SR compared to DT ranges from 0.03[31] to 19.53.[18] Four of the studies were for radiology reports[15,16,31,32] and eight were free text reports.[14,16,19,20,23,24,28,30]

Errors introduced by speech recognition

Document error types in these studies included word omission, word substitution, nonsense phrases, wrong word, punctuation errors, incorrect measurements, missing or added "no". other added words, verb tense, plural, spelling mistakes, or incomplete phrases.[32] Table 2 summarizes error data by class, where available in the studies reviewed, using Kanal's typology.[28] Here, class 0 errors are those that are grammatically correct and do not change meaning. Class 1 errors produce no change in meaning but are grammatically incorrect. Class 2 errors result in a change of text meaning but the error is "obvious." Class 3 errors change meaning in a way that is not obvious to immediate inspection.[28] Four of the five studies reporting error types included class 3 errors, which are likely to be in some way clinically significant with a range of 0.01–0.37 of these occurring per document, including: wrong patient, dose, lab value, and anatomical side errors. Basma's 2011 paper[32] was the only study to provide comparative effectiveness data, with no class 3 errors in the dictation arm, unlike the SR arm which reported a small number of such errors including: wrong word, wrong measurement, and missing or

Table 1: Summary of 23 studies that reported on the performance of speech recognition in a clinical documentation task.

Author	Date	Title	Study Duration (d)	*N*	*n* Reports DT	*n* Reports SR	Accuracy Rate DT (%)	Accuracy Rate SR (%)	Mean Words/ Report DT	Mean Words/ Report SR	Mean Errors/ Report DT	Mean Errors/ Report SR	Mean Dictation Time DT (s)	Mean Dictation Time SR (s)	Mean Editing Time DT (s)	Mean Editing Time SR (s)	Mean TAT DT (min)	Mean TAT SR (min)	*n* Error Classes	SR Technology Utilized	Article SR Disposition
Al-Aynati and Chorneyko	2003	Comparison of voice-automated transcription and human transcription in generating pathology reports	21	7	206	206	99.60	93.60			0.398	6.655			56.80	115.05				IBM Via-Voice Pro 8 with Pathology Vocabulary	+
Basma et al.	2011	Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription	487	33	307	308					0.225	0.698							12	Speech Magic (6.1 SP2, Nuance)	–
Bhan et al.	2008	Effect of voice recognition on radiologist reporting time	14	5	224	244				14.3					83.10	94.20				Powerscribe SR System 4.7 1b, Nuance	+
Chang et al.	2011	Nonclinical errors using voice recognition dictation software for radiology reports: a retrospective audit	183	19		1010						0.477							6	Powerscribe 3.5	Exclude
Chapman et al.	2000	Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department	42	Whole Dept.	400	327											753	133		Dragon Naturally Speaking	+
Chleborad et al.	2013	Evaluation of voice-based data entry to an electronic health record system for dentistry		3										131.79						Dent Voice	+
David et al.	2014	Error rates in physician dictation: quality assurance and medical record production	1		1425	966					0.315								15		–
Hundt et al.	1999	Speech processing in radiology		4	200	200		88.90		135.4			150.00	204.00	66.00		1410	486		Philips SP 6000	+
Ilgner et al.	2006	Free-text data entry by speech recognition software and its impact on clinical routine		1		68		93.15						424.00			3720	1440		IBM ViaVoice, German v10	+
Issenman et al.	2004	Use of voice recognition software in an outpatient pediatric specialty practice	30	1		72	99.97	90.80	259.0	225.0	0.070	19.600	240.00	180.00	17.00	336.00				Dragon Naturally Speaking v6	–
Kanal et al.	2001	Initial evaluation of a continuous speech recognition program for radiology		6		72		89.70											4	IBM MedSpeak v1.1	+
Koivikko et al.	2008	Improvement of report workflow and productivity using speech recognition – a follow-up study	274	30	6037	15558											1486	280		Philips SpeechMagic	+

(continued)

Table 1: Continued.

Author	Date	Title	Study Duration (d)	*N*	*n* Reports DT	*n* Reports SR	Accuracy Rate DT (%)	Accuracy Rate SR (%)	Mean Words/ Report DT	Mean Words/ Report SR	Mean Errors/ Report DT	Mean Errors/ Report SR	Mean Dictation Time DT (s)	Mean Dictation Time SR (s)	Mean Editing Time DT (s)	Mean Editing Time SR (s)	Mean TAT DT (min)	Mean TAT SR (min)	*n* Error Classes	SR Technology Utilized	Article SR Disposition
Krishnaraj et al.	2010	Voice recognition software: effect on radiology report turnaround time at an academic medical center	548	30	149049	156843											1680	762		Talk Technology v3.0	+
McGurk et al.	2008	The effect of voice recognition software on comparative error rates in radiology reports	7	Whole Dept.	727	1160					0.021	0.048							3	Talk Technology V2, DNS	–
Mohr et al.	2003	Speech recognition as a transcription aid: a randomized comparison with standard transcription	28	Whole Dept.	1287	1745							276.00	246.00	1248.00	1176.00				LTI Model-Building Software	–
Pezzullo et al.	2008	Voice recognition dictation: Radiologist as transcriptionist	91	7	100	100	99.90	96.00	225.6	181.7	0.120	4.400	131.00	251.00	10.00	129.00				Talk Technology V2.1.28, AGFA	–
Quint et al.	2008	Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology	84	88	265							0.362							11		–
Ramaswamy et al.	2000	Continuous speech recognition in MR imaging reporting: Advantages, disadvantages, and impact	274	44	4552	5072		92.70	95.0	60.0							5268	2616	3	MedSpeak/ Radiology v1.2	+
Rana et al.	2005	Voice recognition for radiology reporting: is it good enough?	30	4	220	220					0.660	0.890	23.37	31.70						ASR Medispeak	+
Rosenthal et al.	1998	Computers in radiology: computer-based speech recognition as a replacement for medical transcription	14	9	686	656											3737	1459		MedSpeak/ Radiology v1.2	+
Smith et al.	1990	Recognition accuracy with a voice-recognition system designed for anesthesia record keeping	3	31		30		95.30												AARK systems X2	+
Vorbeck et al.	2000	Report generation using digital speech recognition in radiology		2	450	450	99.60	94.50							303.00	253.00	329	275		Philips SP 6000	+
Zemmel and Neil	1996	Status of voicetype dictation for windows for the emergency physician		7		14		91.00												IBM VoiceType Dictation	–

REVIEWS

Figure 2: Decrease in clinical document turnaround time using speech recognition compared to dictation for all studies, plotted by year of study publication.

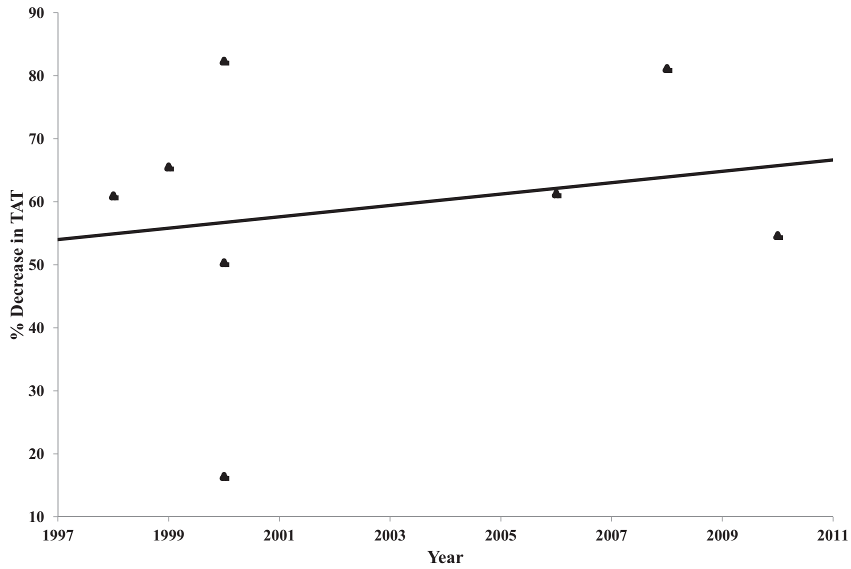
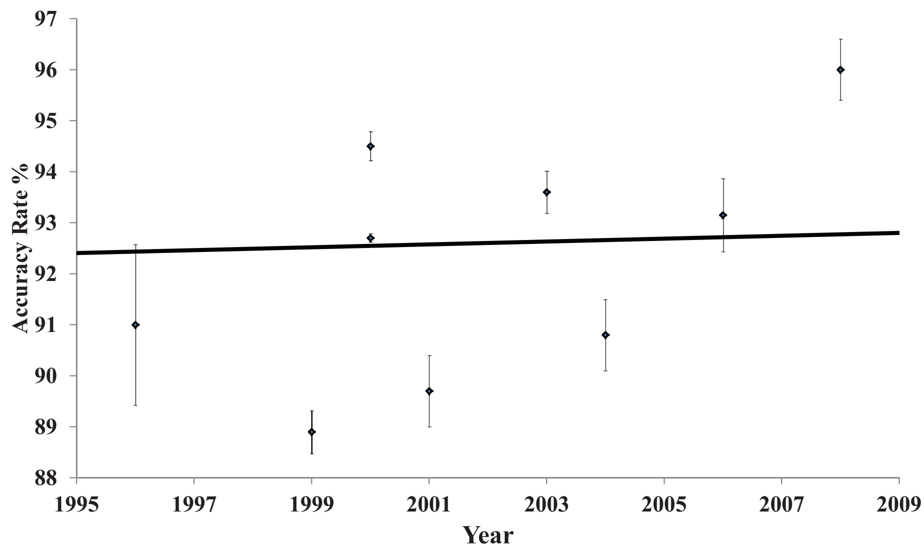


Figure 3: Article speech recognition accuracy rate as a percentage vs year of article publication with 95% confidence intervals.



added "no". No study was designed to test the clinical impact of documentation errors introduced by SR.

Cost–benefit of speech recognition

Some form of economic evaluation was identifiable in seven studies.[15,16,18,20,23,24,28] Four studies found SR-created documentation offered financial benefits,[15,20,23,24] with staff savings of up to £20 000 per year reported in a 2005 study from England. Two US studies in 2004 and 2008 reported financial costs of using SR of up to US$76 250 per annum.[16,18] These cost–benefit studies included: analysis of hardware, software, salaries of both documenting authors, and transcriptionists compared to any efficiencies achieved. The studies excluded any additional start-up costs.

Quality of studies

The studies show significant heterogeneity, with differences in data quality and trial design.

Table 2: Error classes of five studies that reported error types SR and DT (after Kanal, 2001).

Author	Date	Title	SR or DT	Class 0	Errors / Report	Class 1	Errors / Report	Class 2	Errors / Report	Class 3	Errors / Report
				Formatting Errors		Grammatically Incorrect		Meaning Different (error was obvious)		Meaning Different (error not obvious)	
Basma et al.	2011	Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription	SR	Punctuation error	0.16	Verb tense	0.04	Word omission	0.14	Wrong word	0.01
				Spelling mistakes	0.03	Plural	0.06	Word substitution	0.13	Incorrect measurement	0.01
						Incomplete phrase	0.01	Nonsense phrase	0.02	Missing or added "no"	0.01
								Added word	0.15		
			DT	Punctuation error	0.02	Verb tense	0.02	Word omission	0.04	Wrong word	0.00
				Spelling mistakes	0.03	Plural	0.02	Word substitution	0.05	Incorrect measurement	0.00
						Incomplete phrase	0.00	Nonsense phrase	0.00	Missing or added "no"	0.00
								Added word	0.04		
Chang et al.	2011	Nonclinical errors using voice recognition dictation software for radiology reports: a retrospective audit	Combined	Type E – punctuation	0.03	Type F – other including spelling	0.03	Type A – wrong word substitution	0.18	Type C – deletion	0.07
										Type D – insertion	0.13
			SR					Type B – Nonsense phrase	0.02		
			DT					Type B – Nonsense phrase	0.05		
David et al.	2014	Error rates in physician dictation: quality assurance and medical record production	SR			Made up words	0.14	Age mismatch	0.03	Wrong patient	0.07
								Gender mismatch	0.05	Wrong drug name/dosage	0.05
								Wrong name	0.01	Name/dosage	0.04
								Wrong doctor	0.02	Wrong lab values	0.01
								Wrong date	0.01	Left/right discrepancy	0.03
								Other	0.20	Medical discrepancy other	0.00
			DT			Made up words	0.11	Age mismatch	0.04	Wrong patient	0.05
								Gender mismatch	0.02	Wrong drug	0.05
								Wrong name	0.02	Name/dosage	0.01
								Wrong doctor	0.02	Wrong lab values	0.03
								Wrong date	0.03	Left/right discrepancy	0.02
								Other	0.38	Medical discrepancy other	0.04
McGurk et al.	2008	The effect of voice recognition software on comparative error rates in radiology reports	SR			Trivial (no changes to understanding)	0.02	Effect understanding	0.03		
			DT			Trivial (no changes to understanding)	0.01	Effect understanding	0.01		
Ramaswamy et al.	2000	Continuous speech recognition in MR imaging reporting: Advantages, disadvantages, and impact	SR	Irregular spacing	2.98	Spelling errors	0.30			Omissions and duplications	0.372
			DT	Irregular spacing	0.11	Spelling errors	1.12			Omissions and duplications	0.1116

Class 0 errors: no change in meaning and grammatically correct; class 1 errors: no change in meaning but grammatically incorrect; class 2 errors: meaning was different but error was obvious; class 3 errors: meaning was different but not obvious.[28]

Eighteen articles documented study duration, which varied from one day[33] to 548 days.[27] Trial population sizes varied greatly, from one[18,25] to eighty-eight.[34] Three studies looked at "whole departments" where the actual number of subjects in a trial could not be identified. The mean reported population size was seventeen participants per trial.

The number of clinical documents assessed within each study varied fourteen[30] to over three hundred thousand.[27] There were seven articles that categorized errors found within reports,[24,28,31–35] with between three[24,31] and fifteen[33] possible different error categories. The number of different error types studied increased over time from three (1999) to fifteen (2014).

Risk of bias

Each study was assessed for its risk of bias using the Cochrane Collaborations tool for assessing risk of bias. The following risks were identified:

Selection: Participation in the trials varied from compulsory (whole departments) to voluntary enrollment. Where the study was voluntary, it was more likely that those with interest in, and with a positive opinion towards, SR participated.[23] Randomization of report assignment to SR or DT did not always occur. In some cases authors used whichever method they wished to use, or in others a separate facility using SR was compared to another using DT.[32] While the quantity of reports assessed varied greatly, the ratio of SR to DT created reports across all articles was consistently equivalent. Across all studies, the mean number of SR created reports was 11.55% more than those created via DT.

Performance: The studies were composed of a combination of both blind and unblinded trials. In many cases the participants of the study knew which method the report was to be processed by, either SR, DT, or both methods.

Attrition: The mean study duration was 120 days across all studies. The length of the studies and department wide participation led to a turnover of study participants,[27] while others simply faced participant attrition.[18]

Qualitative assessment of speech recognition's value for clinical documentation

The studies were analyzed to determine their overall assessment of the value of SR in support of clinical documentation. Of the reviewed papers, 63.6% were positive about the use of SR and 36.4% were negative. There was a growing consensus among studies published between 1998 and 2001 that SR was of benefit to clinical documentation (Figure 4). However, from 2002 to 2014 there was an even split between positive and negative assessments, suggesting recent studies are more equivocal about SR benefits.

DISCUSSION

SR is a widely used input modality for modern computer devices and has a long pedigree in the clinical setting. Surprisingly, our review revealed that the evidence base documenting the benefits and limitations of SR's use for clinical documentation is limited, incomplete, and relatively neutral to its benefits. Recent studies, which would benefit from more modern SR technologies, are absent.

Medical specific editions of commercial SR packages appeared in the late 1990s with three major players releasing products in 1998 (Dragon NaturallySpeaking with Medical add-on, IBM ViaVoice 98 with General Medicine Vocabulary, and L&H Voice Xpress for medicine, General Medicine Edition). The overwhelming enthusiasm for SR in the infancy of its commercial release in the 1990s is reflected in the views of the studies and editorials of that period.[29,22,36,37] That enthusiasm has moderated over time to a more balanced view of SR, recognizing both its benefits and limitations.

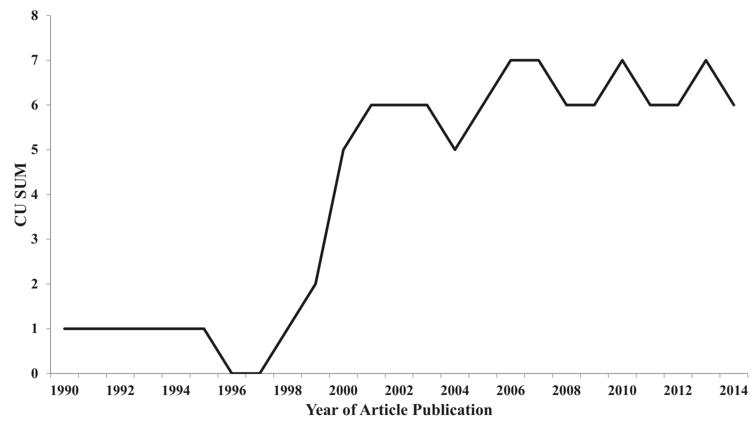
There were clear overall system-level benefits in relation to documentation speed when using SR, with dramatic reductions reported for overall TAT for report creation. This is mainly due to the virtually instant delivery of reports possible with SR based systems. This improvement hides an editing and document creation time cost that falls directly on the clinician. The effective clinical adoption of technologies often depends on local costs being offset by local benefits, and the relatively low uptake of SR to date might in part be due to an imbalance in cost over benefit for the clinician preparing reports.

A modest increase in the accuracy of SR over time has been reported within these studies. There are numerous technological reasons for this including: improved quality of microphones, SR software packages, and underlying computer hardware. In fact many SR software developers now claim accuracy rates of up to 99%.[38] However, high accuracy rates do not necessarily mean that SR is clinically safe, and several studies have reported a range of errors, some of which are clinically significant and could lead to patient harm. Reported errors included: creating documentation for the wrong patient, wrong drug name or dosage, wrong lab values, left/right anatomical discrepancy, medical discrepancy, age or gender mismatch, wrong doctor name, wrong date, made up words and acronyms, irregular spacing, spelling errors, omissions, or duplications.[24,31,33] Lower rates for clinically significant errors were reported for DT, which was partially due to documentation being completed by skilled and practiced transcriptionists who offered an additional safety check on the content of a clinical document.

There are numerous variables that potentially affect SR system performance that were not adequately captured within the studies. These variables include: user training and experience, SR speed, microphone quality, author accent or speech impediment, dictation interruptions, background noise level, and other environmental conditions. In the absence of such data, there is the need to cautiously generalize the performance reported in these studies to expected real world performance. In other words, good performance under controlled conditions may not be replicated in clinical settings.

Based on these studies alone, it is impossible to establish whether SR is an efficient or effective input modality for the creation of clinical documentation and for which settings and modes of use it is best suited. The widespread adoption of SR,

Figure 4: Cumulative sum of expressed disposition of studies (positive/negative) towards the use of speech recognition for clinical documentation (1990–2014). (Each positive article adds +1 to the cumulative score; each negative article reduces the cumulative score by –1).



however, should provide evidence that such benefits exist, but additional research is needed in a number of areas before this is robustly established:

Impact on clinical processes and outcomes: No studies in this review looked at the impact of SR on clinical outcomes. While the creation of electronic documentation for the electronic health record is more likely to see improvements in organizational process rather than clinical outcomes,[39] SR use may result in changes to the size and content of documents, with the potential to impact clinical decisions. The introduction of SR will also result in significant changes to the business processes undertaken by an organization and these changes along with their follow-on effects may need to be analyzed.

Impact on clinicians: The studies in this review revealed that system level benefits masked clinician costs. While SR should intuitively be easier than more traditional input modalities, it appears to come with a time cost when editing documents. If the use of SR also increases cognitive load for clinicians, it may have an impact on other clinical tasks, time efficiency, and error rates.[40]

Impact on patient safety: The introduction of information technologies are typically associated with some risks to patients.[41] Technology both creates new error classes, as well as new opportunities for user errors, and these are likely to vary with the specific technology, its users, their setting, training, and tasks. Given the clear evidence of new errors associated with SR, any evaluation of benefits requires a diligent assessment of possible harm.

Comparative effectiveness: This review contained no studies that compared SR with the current dominant input paradigm of keyboard and mouse, a reflection perhaps of the significant role that dictation plays in settings such as radiology where most studies were conducted. Any definitive assessment of SR will need to occur in relation to common alternatives including keyboard input.

Effectiveness for nondocumentation tasks: SR may have great potential for a variety of tasks such as: order entry, alert management, and patient handoffs. However, these aspects were not addressed throughout the majority of the studies. Many such tasks will require smaller controlled vocabularies, making the task potentially easier, but perhaps would suffer because they would be enacted in noisier environments.

Alternate input platforms: Clinical practice uses a wide variety of information technology platforms, from traditional computer workstations, to smart phones, tablet devices, and wearable devices such as head-mounted displays and glasses, all of which might use SR in different ways. The utility of SR is likely to vary across these platforms.

Limitations of this review

Heterogeneity across study task, design, and population, as well as low sample sizes, precluded subgroup analyses. While this allowed some analysis of temporal trends, it prevents direct comparison of many studies. Similarly, the sophistication of study designs increased with time, and early studies were often not of comparable quality to more recent papers. Other limitations include scope, timespan, and economic variables.

Scope: There were few comparative effectiveness assessments of SR. When comparisons were made, they were with DT. Most studies were focused on the preparation of radiological reports, and few examined the creation of clinical documentation that would appear in the main sections of a patient record. The use of voice for advanced functions such as navigation and control were not explored. Many variables that could affect real-world performance were not explored or reported.

Timespan: The studies occurred over more than twenty years, and covered many generations of SR technology. SR systems used included IBM MedSpeak and ViaVoice[20,22,24,25,28,30]; Philips – SP6000, SpeechMagic[14,19,26]; Nuance's Dragon Naturally Speaking, Powerscribe and

REVIEWS

SpeechMagic 2009+[18,21,23,32,35]; AGFA's Talk Technology[16,27,31]; and AARK Systems, ASR Medispeak, the Dent Voice Prototype, and LTI.[15,17,29,42]

Economic benefit: Cost–benefit analyses were limited and poorly described in the studies reviewed, and most were not directly comparable given the significant variations in clinical setting, technology, installation, maintenance, and support costs.

CONCLUSION

This review reveals that SR is a potentially valuable tool for clinical documentation. However, any advantages must be weighed against the potential for time penalties for clinicians, the potential for new errors, and unclear cost–benefits in some clinical settings. The research evidence is surprisingly sparse, and there remain many unanswered questions and unexplored opportunities. While SR may not be viable for all clinicians in all scenarios, it is currently not possible to clearly articulate the tasks and clinical settings in which its use is clearly of benefit, and where it should perhaps be avoided. New and emerging platforms including smartphones and wearable devices such as head-mounted displays and glasses seem ideally suited to the clinical use of SR, making this an area deserving of much greater research attention.

FUNDING

This work was supported by the NHMRC Centre for Research Excellence in eHealth (APP1032664).

COMPETING INTERESTS

The authors declare that they have no competing interests.

CONTRIBUTORS

T.H. and E.C. conceived the study and its design. T.H. conducted the research, the primary analysis, and the initial drafting of the paper. E.C. contributed to the analysis and drafting of the paper and both T.H. and E.C. approved the final manuscript. T.H. is the corresponding author.

ACKNOWLEDGEMENTS

Diana Arachi was the second reviewer who assisted in the phase of article inclusion and data extraction.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Johnson M, Lapkin S, Long V, *et al*. A systematic review of speech recognition technology in health care. *BMC Med Inform Decis Mak.* 2014; 14(1):94.
2. Herman SJ. Speech recognition and the creation of radiology reports. *Appl Radiol.* 2004;33(5):23–28.
3. Lawrence D. Can you hear me now? Voice recognition for the EMR has made big strides, and many say meaningful use requirements will accelerate adoption. Healthcare informatics: the business magazine for information and communication systems. 2009;26(12).
4. Neustein A, Markowitz JA. *Mobile Speech and Advanced Natural Language Solutions*. New York: Springer; 2013.
5. Bliss MF. *Speech Recognition for the Health Professions: (using Dragon NaturallySpeaking)*. Upper Saddle River, N.J.: Prentice Hall; 2005.
6. Madisetti V. *Video, Speech, and Audio Signal Processing and Associated Standards*. Boca Raton, FL: CRC Press; 2009.
7. Paulett JM, Langlotz CP. Improving language models for radiology speech recognition. *J Biomed Inform.* 2009;42(1):53–58.
8. Gales M, Young S. The application of hidden Markov models in speech recognition. *Found Trends Signal Process.* 2008;1(3):195–304.
9. Eddy SR. What is a hidden Markov model? *Nat Biotech*. 2004;22(10):1315–1316.
10. Indurkhya N, Damerau FJ. *Handbook of Natural Language Processing*. Boca Raton, FL: CRC Press; 2012.
11. Lafferty J, McCallum A, Pereira FC. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 2001.
12. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Int Med.* 2009;151(4):264–269.
13. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull.* 1968;70(4):213.
14. Hundt WS, Scharnberg O, Hold B, *et al*. Speech processing in radiology. *Eur Radiol.* 1999;9(7):1451–1456.
15. Rana DSH, Shepstone G, Pilling L, Cockburn J, Crawford, M. Voice recognition for radiology reporting: is it good enough? *Clin Radiol*. 2005;60(11):1205–1212.
16. Pezzullo JA, Tung GA, Rogg JM, *et al*. Voice recognition dictation: Radiologist as transcriptionist. *J Digital Imag.* 2008;21(4):384–389.
17. Mohr DNT, Turner DW, Pond GR, *et al*. Speech recognition as a transcription aid: a randomized comparison with standard transcription. *JAMIA.* 2003;10(1):85–93.
18. Issenman RM, Jaffer IH. Use of voice recognition software in an outpatient pediatric specialty practice. *Pediatrics*. 2004;114(3):e290-e293.
19. Vorbeck F, Ba-Ssalamah A, Kettenbach J, Huebsch P. Report generation using digital speech recognition in radiology. *Eur Radiol.* 2000;10(12):1976–1982.
20. Al-Aynati MM, Chorneyko KA. Comparison of voice-automated transcription and human transcription in generating pathology reports. *Arch Pathol Lab Med.* 2003;127(6):721–725.
21. Bhan SN, Coblentz CL, Norman GR, Ali SH. Effect of voice recognition on radiologist reporting time. *Can Assoc Radiol J.* 2008;59(4):203–209.
22. Rosenthal DIC, Dupuy FS, Kattapuram DE, *et al*. Computers in radiology: computer-based speech recognition as a replacement for medical transcription. *Am J Roentgenol.* 1998;170(1):23–25.
23. Chapman WW, Aronsky D, Fiszman M, Haug PJ. Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department. *Proc/AMIA Ann Symp AMIA Symp.* 2000:131–135.
24. Ramaswamy MR, Chaljub G, Esch O, Fanning DD, VanSonnenberg E. Continuous speech recognition in MR imaging reporting: advantages, disadvantages, and impact. *Am J Roentgenol.* 2000;174(3):617–622.
25. Ilgner J, Duwel P, Westhofen M. Free-text data entry by speech recognition software and its impact on clinical routine. *Ear, Nose Throat J.* 2006;85(8):523–527.
26. Koivikko MP, Kauppinen T, Ahovuo J. Improvement of report workflow and productivity using speech recognition—a follow-up study. *J Digit Imaging.* 2008;21(4):378–382.
27. Krishnaraj A, Lee JK, Laws SA, Crawford TJ. Voice recognition software: effect on radiology report turnaround time at an academic medical center. *Am J Roentgenol.* 2010;195(1):194–197.

28. Kanal KM, Hangiandreou NJ, Sykes AM, *et al*. Initial evaluation of a continuous speech recognition program for radiology. *J Digit Imaging.* 2001;14(1):30–37.
29. Smith NT, Brien RA, Pettus DC, Jones BR, Quinn ML, Sarnat A. Recognition accuracy with a voice-recognition system designed for anesthesia record keeping. *J Clin Monitor.* 1990;6(4):299–306.
30. Zemmel NJ, Park SM, Schweitzer J, O'Keefe JS, Laughon MM, Edlich RF. Status of voicetype dictation for windows for the emergency physician. *J Emerg Med.* 1996;14(4):511–515.
31. McGurk S, Brauer K, Macfarlane TV, Duncan KA. The effect of voice recognition software on comparative error rates in radiology reports. *Brit J Radiol.* 2008;81(970):767–770.
32. Basma S, Lord B, Jacks LM, Rizk M, Scaranelo AM. Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription. *Am J Roentgenol.* 2011;197(4):923–927.
33. David GC, Chand D, Sankaranarayanan B. Error rates in physician dictation: quality assurance and medical record production. *Int J Health Care Qual Assur.* 2014;27(2):99–110.
34. Quint LE, Quint DJ, Myles JD. Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. *J Am Coll Radiol.* 2008;5(12):1196–1199.
35. Chang CA, Strahan R, Jolley D. Non-clinical errors using voice recognition dictation software for radiology reports: a retrospective audit. *J Digit Imaging.* 2011;24(4):724–728.
36. Belton K, Dick R. Voice-recognition technology: key to the computer-based patient record. *J Am Med Record Assoc.* 1991;62(7):27–32, 36, 38.
37. Clark S. Implementation of voice recognition technology at provenant health partners. *J Am Health Inform Manag Assoc.* 1994;65(2):34, 36, 38.
38. Nuance Communications. *Dragon NaturallySpeaking 13 Premium Data Sheet - Nuance Communications*. 2014:2.
39. Coiera E. *Guide to Health Informatics*. 3rd edn. Boca Raton, FL: CRC Press; 2015.
40. Coiera E. Technology, cognition and error. *BMJ Qual Saf* 2015;24(7):417–422.
41. Coiera E, Aarts J, Kulikowski C. The dangerous decade. *JAMIA.* 2012; 19(1):2–5.
42. Chleborad KD, Zvara T, Hippmann K, *et al*. Evaluation of voice-based data entry to an electronic health record system for dentistry. *Biocybernetics Biomed Eng.* 2013;33(4):204–210.

AUTHOR AFFILIATION

Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, L6 75 Talavera Rd, North Ryde, NSW 2109, Australia

REVIEWS